Appendix X

Information on the Sacramento County Regression Equation

Clinton P. MacDonald

Daniel M. Alrick

Dianne S. Miller

1. Introduction

Ozone predictions from regression equations were used by SMAQMD to estimate wildland fire's contribution to peak 1-hr ozone concentrations in Sacramento, California. This appendix provides background information about regression equations (Section 2), how Sonoma Technology, Inc. (STI) developed the regression equation used for the exceptional event analysis (Section 3), and how the equation was used to calculate the wildland fire's contribution to ozone in Sacramento County (Section 4). The wildland fire's contribution to peak 1-hr ozone concentrations is discussed in the main body of the report.

2. Background on Regression Equations

Regression is a statistical method for describing relationships among variables. For estimating air quality concentrations, regression equations are developed to describe the relationship between pollutant concentrations (referred to as the predictand, what is being predicted) and primarily meteorological variables (referred to as the predictors). Regression equations have been successfully used to predict daily or subdaily pollutant concentrations in many areas of the country (Cassmassi, 1987; Hubbard and Cobourn, 1997; Ryan, 1994; Dye et al., 1996; U.S. Environmental Protection Agency, 2003).

Because regression equations are developed with several years of data, they represent the relationship between air quality and meteorology under typical emission patterns. Therefore, the difference between the predictions and observations can provide a reasonable estimate of the air pollution caused by unusual emissions (e.g., emissions from wildland fires).

How Regression Equations Work

If two variables are correlated, a mathematical equation can generate a line or a curve that depicts the relationship between those variables. With this mathematical equation, one variable (e.g., ozone) can be predicted from other variables (e.g., meteorology). Multi-linear regression is most commonly used to predict ozone (Equation 1) from multiple variables. However, curvilinear regression (Equation 2) is also useful in predicting ozone because it captures the non-linear relationships of ozone and predictor variables.

Ozone =
$$c_1 V_1 + c_2 V_2 \dots c_n V_n + constant$$
 (1)

Ozone =
$$c_1 V_1 + c_2 V_2^2 + c_3 V_3^3 \dots, c_n V_n^n + constant$$
 (2)

where:

Ozone = predictand

c = coefficients (weighting factors)
V = meteorological predictor variables

An example of a multi-linear regression equation is shown in **Equation 3**. This regression model was developed for forecasting hourly ozone concentrations for Sacramento, California. The variables are described in **Table 1**. To predict ozone concentration on a given day, one simply inputs the value of each meteorological variable for that day into the equation.

1-hr Ozone = $\exp (13.72 - 0.03 \text{*Clouds} - 0.04 \text{*WindSpeed1} + 0.01 \text{*WindSpeed2})$

- + 0.0002*WindDirection 0.01*Pressure 0.02*DewPoint
- + 0.03*AloftTemperature 0.009*AloftWindSpeed
- + 0.009*TemperatureDifference)

(3)

Table 1. Variables used in regression Equation 3. The surface meteorological conditions are for Sacramento, California. The aloft meteorological conditions are for Oakland, California.

Variable Abbreviations	Description		
Clouds	Average hourly cloud cover from 6:00 a.m. PST to 6:00 p.m. PST where clear = 0, partly cloudy = 1, mostly cloudy = 2, and overcast = 3		
WindSpeed1	Average wind speed from 6:00 a.m. PST to 12:00 p.m. PST in m/s		
WindSpeed2	Surface wind speed at 00Z (4:00 p.m. PST on the previous day) in m/s		
WindDirection	Surface wind direction at 00Z (4:00 p.m. PST on the previous day)		
Pressure	Surface pressure at 12Z (4:00 a.m. PST) in mb		
DewPoint	Surface dew point temperature at 12Z (4:00 a.m. PST) in °C		
AloftTemperature	925-mb temperature at 00Z (4:00 p.m. PST on the previous day) in °C		
AloftWindSpeed	925-mb wind speed at 12Z (4:00 a.m. PST) in m/s		
TemperatureDifference	Temperature difference from 850 mb to the surface at 12Z (4:00 a.m. PST) in °C		

3. Regression Equation Development

As part of STI's Forecasting and Outreach Support project with the Sacramento Metropolitan Air Quality Management District (SMAQMD), in 2004 STI developed a regression equation to assist with daily ozone forecasting for the Sacramento region. To develop the regression equation, STI staff completed the following tasks:

Data Processing

- Obtained and processed six years (1997-2003) of May through October ozone data from the U.S. Environmental Protection Agency (EPA) for Sacramento County. The sites include Sacramento T Street, Del Paso Manor, Airport Blvd., North Highlands, Elk Grove, Sloughhouse, and Folsom.
- Processed six years (1997-2003) of surface and upper-air meteorological data from the National Oceanic and Atmospheric Administration (NOAA) for the Sacramento area.

- Performed general quality control of the air quality and meteorological data.
- Populated a Microsoft Access database with Sacramento air quality and meteorological data, linked by date. New variables, such as average wind speeds, average cloud cover, minimum and maximum temperature, and stability parameters, were computed from the surface and upper-air meteorological data that capture meteorological processes important to air quality.

Regression Equation Development

- Imported final data table from Microsoft Access into SYSTAT statistical software.
- Determined proper software inputs for the regression algorithm to yield statistically sound equations. The software inputs included:
 - Tolerance limits the similarity or collinearity of input variables. For example, 950-mb temperature and 925-mb temperature are closely related; therefore, only one factor will be allowed into the equation. STI used tolerance values between 0.3 and 0.5.
 - Estimation controls the method used to enter and remove variables from the equation. STI used Stepwise estimation in which variables are entered or removed from the model one at a time.
 - Stepwise options control the entry and removal of variables. STI used the Forward and Automatic options:
 - Forward begins with no variables in the model. At each step, SYSTAT adds the variable with the smallest Enter value.
 - Automatic SYSTAT automatically adds a variable to the model at each step for the Forward option.
 - Probability specifies probabilities to enter and remove a variable from the model. A variable is entered into the model if its alpha value is less than the specified Enter value, and it is removed from the model if its alpha value is greater than the specified Remove value. STI used probabilities between 0.05 and 0.15.
 - MaxStep dictates the maximum number of steps (variables) allowed. STI used a maximum of 10 variables.
 - Force forces the first n variables listed to remain in the equation. STI did not force any variables into the equations (i.e., n = 0).
- Evaluated the output variables for physical sense (i.e., higher ozone expected to be predicted with lower wind speeds).
- Evaluated the statistical strength of the equation. Output metrics included
 - Standard error shows the average deviation of a sample from the expected mean; small error is best.
 - T-test shows whether each variable in the equation is statistically significant by comparing the mean of the predictions with and without the new variable.
 - P2-tail shows the likelihood of the equation's performing well with a random sample, such as when used operationally Values less than 0.05 are desired, indicating a 95% probability that the model will work with a random sample.

 Standard coefficient of a variable – shows how much the equation output would change by altering the variable by one standard deviation. In other words, it shows how heavily each variable contributes to the overall equation.

Regression Equation Testing

- Compared the regression output to observations in order to evaluate accuracy.
 Metrics used for evaluation included
 - accuracy based on Air Quality Index (AQI) category;
 - accuracy based on predicted concentration;
 - bias; and
 - average error.
- Tested regression equations on independent data sets using the same metrics listed above.
- Selected the best equation to use, basing the selection on test results.

Test Results Associated with Equation Development

The final equation used to predict daily peak 1-hr maximum ozone concentration for Sacramento County is shown in Equation 3. Immediately after developing the equation, we evaluated the equation's performance by comparing the predictions from the equations to the observations from both the data set used to develop the equations and a data set reserved for testing only. The results for the correlation coefficient (r^2) and bias values are shown in **Table 2**. The overall performance is good and is consistent with performance from other regression equations developed for air quality forecasting programs (U.S. Environmental Protection Agency, 2003).

Table 2. Sacramento County 1-hr ozone equation test results.

Metric	Result
Development data set r ²	0.75
Development data set bias	15.2 ppb
Independent data set r ²	0.58
Independent data set bias	8 ppb

4. Application to Exceptional Event Analysis

To estimate the wildland fire contribution to ozone concentrations in Sacramento, the following general steps were performed:

 STI used the Sacramento regression equation to calculate daily peak 1-hr ozone predictions for Sacramento County for each day during May through October 2007. The meteorological input data used in the regression equation were the output data from the National Center for Environmental Prediction's (NCEP) Eta model. The Eta model is a weather forecast model that was used by the National Weather Service to produce daily weather forecasts until it was replaced by the Weather Research and Forecast model. Several variations of the Eta model predictions were available. STI used three:

- A. The predictions produced by the 12:00 UTC model run, which provided the most recent meteorological predictions each day to run the regression equations. We will refer to the ozone predictions that used this input data as the "12 Prediction."
- B. The predictions produced by the 12:00 UTC model run but adjusted by NCEP using statistical techniques called Model Output Statistics (MOS). We will refer to the ozone predictions that used this input data as the "12 MOS Prediction."
- C. The predictions produced by the 00:00 UTC run, which provided meteorological predictions that are 12 hours less recent than 12:00 UTC model predictions. We will refer to the ozone predictions that used this input data as the "00 Prediction."

Eta model weather data, rather than observed data, were used to predict ozone because the model did not account for the wildfire's influence on meteorology; thus, the Eta predictions represent meteorological conditions (and ultimately ozone via the regression equations) that would be expected in the absence of the wildfires.

- 2. STI calculated the average daily differences (bias) between the daily ozone predictions and the observed ozone concentrations on all non-smoke impact days in 2007. The average bias between the predicted and observed ozone was 8.4, 8.6, and +13.3 ppb for the 12 Prediction, 12 MOS Prediction, and 00 Prediction, respectively. These positive biases mean that, on average, the regression equation overestimated ozone concentrations. The positive bias is likely due to reductions in ozone precursor emissions between the period for which the equations were developed (1997-2003) and the years that they were applied (2007 and 2008). This result was expected.
- 3. To correct for the positive biases, to make these equations more accurate for 2007 and 2008, STI subtracted the average biases (8.4, 8.6, and +13.3) from the associated daily predictions (12 Prediction, 12 MOS Prediction, and 00 Prediction) for May through October of 2007 and 2008.
- 4. SMAQMD used the "bias-adjusted" predictions from Step 3 to estimate the wildfire contribution to ozone. In particular, for June 23, June 27, July 7, and July 10, 2008, SMAQMD subtracted the "bias-adjusted" 1-hr ozone predictions for Sacramento County (including the 12 Prediction, 12 MOS Prediction, and 00 Prediction) from the peak observed 1-hr ozone concentration in Sacramento County. The results were three estimates of wildfire contribution to ozone in Sacramento County for June 23, June 27, July 7, and July 10, 2008. These results are presented in the main body of the report.
- 5. STI estimated the error of the 12 Prediction, 12 MOS Prediction, and 00 Prediction methods. In particular, for each prediction method, STI
 - A. Calculated the difference between the daily bias-adjusted ozone prediction and daily observed ozone concentrations for May through

- October for the years 2006, 2007, 2008, except for the fire impact days of 2008.
- B. Plotted the distribution of error and determined that the errors are "normally distributed" about the mean; i.e., the distribution is symmetric. This indicates that there is no directional bias and the errors are random. The distributions for the 12 Prediction, 12 MOS Prediction, and 00 Prediction are shown in **Figures 1, 2, and 3,** respectively. As indicated by each of the three histograms, the majority of days had relatively small prediction errors, while very few days had large prediction errors.
- C. Calculated the value at which 68% and 95% of the positive daily errors, and 68% and 95% of the negative daily errors, fall within the mean error for each prediction method. For a normal distribution of data, the 68th and 95th percentiles correspond to roughly one and two standard deviations from the mean, respectively. The use of the 95th percentile (approximately two standard deviations) provides a conservative upper and lower bounds of the predictions. The 12 Prediction, 12 MOS Prediction, and 00 Prediction 95% error values are shown in **Table 3**.

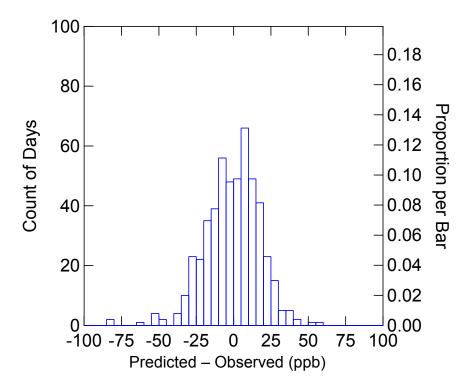


Figure 1. Distribution of ozone prediction error using meteorological input data from the 12:00 UTC ETA model run (12 Prediction).

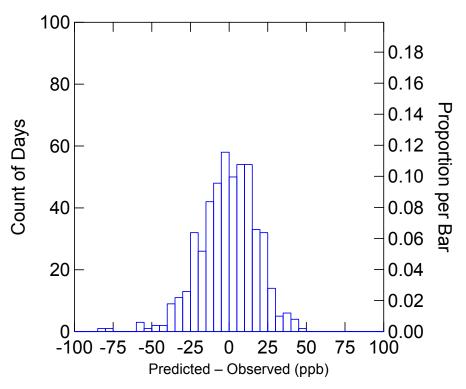


Figure 2. Distribution of ozone prediction error using meteorological input data from the 12:00 UTC ETA MOS model run (12 MOS Prediction).

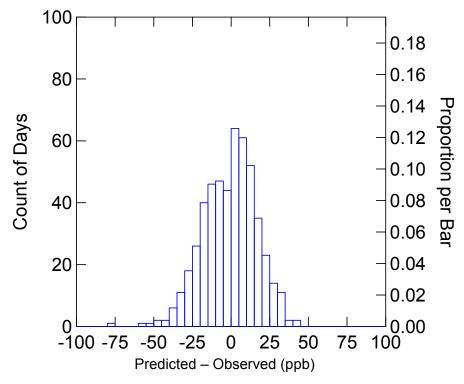


Figure 3. Distribution of ozone prediction error using meteorological input data from the 00:00 UTC ETA model run (00 Prediction).

Table 3. Summary of error values.

	12 Prediction	12 MOS Prediction	00 Prediction
Error value (ppb) at which 95% of positive errors are within the mean	+32	+33	+31
Error value (ppb) at which 68% of positive errors are within the mean	+16	+17	+15
Error value (ppb) at which 68% of negative errors are within the mean	-18	-19	-18
Error value (ppb) at which 95% of negative errors are within the mean	-35	-39	-35

5. References

- Cassmassi J.C. (1987) Development of an objective ozone forecast model for the South Coast Air Basin. Presented at the *Air Pollution Control Association 80th Annual Meeting, New York, NY, June 21-26.*
- Dye T.S., Ray S.E., Lindsey C.G., Arthur M., and Chinkin L.R. (1996) Summary of ozone forecasting and equation development for the air districts of Sacramento, Yolo-Solano, and Placer. Vol. I: ozone forecasting. Vol. II: equation development. Final report prepared for Sacramento Metropolitan Air Quality Management District, Sacramento, CA, by Sonoma Technology, Inc., Santa Rosa, CA, STI-996210-1701-FR, December.
- Hubbard M.C. and Cobourn W.G. (1997) Development of a regression model to forecast ground-level ozone concentration in Louisville, KY. *Atmos. Environ.* (submitted).
- Ryan W.F. (1994) Forecasting severe ozone episodes in the Baltimore metropolitan area. *Atmos. Environ.* **29**, 2387-2398.
- U.S. Environmental Protection Agency (2003) Guidelines for developing an air quality (ozone and PM_{2.5}) forecasting program. Prepared by Dye T.S., MacDonald C.P., Anderson C.B., Hafner H.R., Wheeler N.J.M., and Chan A.C. at Sonoma Technology, Inc., Petaluma, CA, STI-902461-2295-FR, EPA-456/R-03-002, June. Available on the Internet at http://www.epa.gov/airnow/aq forecasting guidance-1016.pdf.

Addendum to Appendix X

Primary documentation of conditions connecting several high ozone events with unusual wildfires has already been developed. In response to a request from U.S. EPA for additional information on certain statistical methods, Sonoma Technology, Inc (STI) prepared a document titled "Appendix X – Information on the Sacramento County Regression Equation". The analysis presented here is an addendum to the "Appendix X" document.

The analysis presented below provides substantial additional support for the conclusion that in the absence of the extensive and persistent wildfires in 2008, ozone values measured on June 23, 2008, June 27, 2008, and July 10, 2008 in Sacramento County, California would not have exceeded the federal 1-hour ozone standard

Data

This analysis makes use of model outputs and observed concentrations described previously in Appendix X. The data include daily maximum 1-hour ozone values in Sacramento County, and daily predicted ozone values produced by Sonoma Technology, Inc. (STI). STI provided four different daily predicted values for ozone, each the product of a different prediction model. The predicted values from the model identified as "00Z Eta" were used in this analysis. The methods used for this addendum were designed to compensate naturally for multiple types of biases that may differ between the alternative models, so the conclusions would likely be the same if a different prediction model had been selected (details on page 3).

Analysis

The data are summarized and presented in two different ways to provide additional support concerning two vital questions:

- Question #1: Is there clear evidence that the extensive and persistent wildfires in California in 2008 caused a net increase in measured ozone levels?
- Question #2: Is there strong evidence that several exceedances of the Federal 1-hour ozone standard would not have occurred "but for" the effect of wildfires on ozone levels.

Question #1: Is there clear evidence that the extensive and persistent wildfires in California in 2008 caused a net increase in measured ozone levels?

Answer: Yes.

Ozone levels increase and decrease from one day to the next throughout the ozone season. These everyday ups and downs are caused by regular changes in meteorological conditions and in emissions of ozone precursors. Figure 1 shows the regular, everyday behavior (typical variability) of daily maximum 1-hour ozone values (solid triangles) in Sacramento County from May 1 through September 30, 2007, a season that was largely unaffected by wildfires. Everyday ups and downs make it hard to identify exceptional ups and downs, the ones caused by irregular meteorology and/or emissions.

Figure 1 also shows predicted values for the daily observed maximum 1-hour ozone values (open diamonds) in Sacramento County. The predicted values represent the expected value for ozone based on forecasted meteorological conditions, so the ups and downs in these values tend to track the regular effects of weather on daily ozone.

Similar to Figure 1, Figure 2 shows the observed and predicted values for daily maximum 1-hour ozone in Sacramento County for 2008. Highlighted in figure 2 are five dates on which measured ozone exceeded the Federal 1-hour standard for ozone. Further analysis provides strong evidence that these exceedances are the result of irregular (fires) rather than regular causes.

Figures 1 and 2 showed that the daily predicted ozone tends to rise and fall along with the observed ozone. By subtracting the predicted values from the observed values, the daily differences will have a "zero" baseline. Some days will have a positive difference because the measured ozone was above the predicted ozone, and some days will have a negative difference because the measured ozone was below the predicted ozone.

Figures 4, 5, 6, 7, 8, and 9 track the daily differences between observed ozone and predicted ozone by month. In these figures, irregular ups and downs in daily ozone stand out from the regular ups and downs. To help in this regard, the figures include a horizontal dashed line representing a "Regular Upper Limit" which is exceeded by regular differences only 5% of the time, and even then the excursions are usually small. The value for the line is the difference (observed minus predicted) that separates the highest 5% from the lower 95% of regular daily differences. Figure 3 shows how the value of the line (27.6 ppb) was determined. A corresponding lower limit is not shown, as the appropriate focus is on high values that are potential exceptional events. Unlike regular differences,

irregular differences stand out because the excursions are large or because there are many more than the 5% that are expected for regular days.

Figure 3, shows a histogram of the daily differences (observed ozone minus predicted ozone). The histogram represents regular differences because the data do not include June 23 to July 31, 2008, the period most strongly affected by wildfires in 2008. August 2008 was also not included because that month is considered borderline with respect to effects of lingering fires. Figure 3 is annotated showing that 95% of the daily differences did not exceed 27.6 ppb ozone. So, the "Regular Upper Limit" line in Figures 4 – 9 has a constant value of 27.6 ppb.

The use of a percentile cut-point (95% point) in these analyses naturally compensates for a variety of possible biases that could differ between the alternative predictive models. An overall additive bias will be built in to each of the daily differences, and that bias would become a part of every percentile drawn from the set of differences. Biases at the high end are unaffected by possibly different biases at the low end, because the distribution of percentiles does not use parametric assumptions, such as "normality". The percentiles follow the data-driven shape of the distribution.

Figure 4 shows data for the month of May in 2007 and 2008 and characterizes regular variability in the daily differences (observed ozone minus predicted ozone).

Figures 5 and 6 show data for June and July in 2007 and 2008. The 2007 data exhibit regular differences, the typical behavior of daily maximum 1-hour ozone. The 2008 data show regular differences through June 22, but then the pattern changes. From June 23, 2008 through July 26, 2008, sixteen of the thirty-four daily differences exceed the regular upper limit, and some of the differences are quite large. Instead of the regularly expected 5%, the figures show a highly irregular 47% (16 of 34) of daily differences above the Regular Upper Limit. During this highly irregular period, northern California experienced extensive and pervasive wildfires. The possibility of extreme meteorology is not reasonable, since the daily differences are already "normalized" with respect to meteorological conditions known to affect ozone levels in the Sacramento area.

Figures 7 and 8 show data for August and September in 2007 and 2008. In these figures, the behavior of the daily differences is mostly regular, though August 2008 is considered borderline with respect to possible effects of lingering wildfires.

This analysis presents very strong statistical evidence that ozone levels in the Sacramento area on many days from June 23, 2008, through July 26, 2008, were enhanced by the well-documented wildfires that were pervasive and persistent in northern California during that time.

Question #2: Is there strong evidence that several exceedances of the Federal 1-hour ozone standard would not have occurred "but for" the effect of wildfires on ozone levels?

Answer: Yes.

The preceding analysis established a causal connection between wildfires and high ozone levels in the Sacramento area from June 23 through July 26, 2008. The following analysis will show that ozone levels on four of five specific dates would not have exceeded the level of the Federal ozone standard but for the effects of the wildfires.

Figures 10, 11, 12, 13, and 14 track daily observed ozone together with a "day-specific regular upper limit". Each day-specific limit is that day's predicted ozone plus the "regular upper limit" (27.5 ppb) used in the previous analysis.

When an observed ozone value exceeds the day-specific regular upper limit, it is likely that some irregular cause contributed to the measured value. When an irregular event known to enhance ozone formation, such as the 2008 wildfires, coincides with an exceedance of the day-specific limit, the convergence of the two events provides strong evidence that the exceedence would not have occurred but for irregular event, in this case the wildfires.

During the wildfire period, observed ozone sometimes exceeds the federal standard when the day-specific upper limit does not exceed the federal standard. For those days, the evidence is strong that observed ozone would not have exceeded the standard "but for" the effects of the wildfires. Figures 11 and 12 show that these conditions prevailed on June 23, June 27, July 10, and July 25, 2008. Table 1 summarizes these results.

Figure 1: Sacramento County: Comparison of 00Z Eta (corrected) and Observed Daily Max. 1-hr Ozone Concentration for May - Sep., 2007

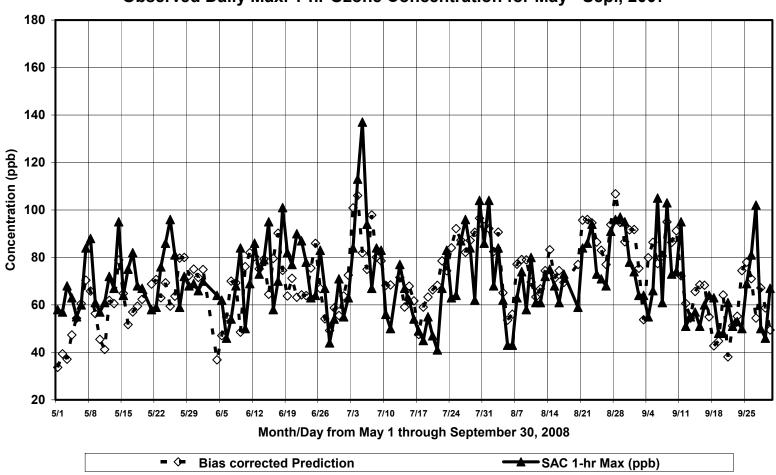


Figure 2: Sacramento County: Comparison of 00Z Eta (corrected) and Observed Daily Max. 1-hr Ozone Concentration for May - Sep., 2008

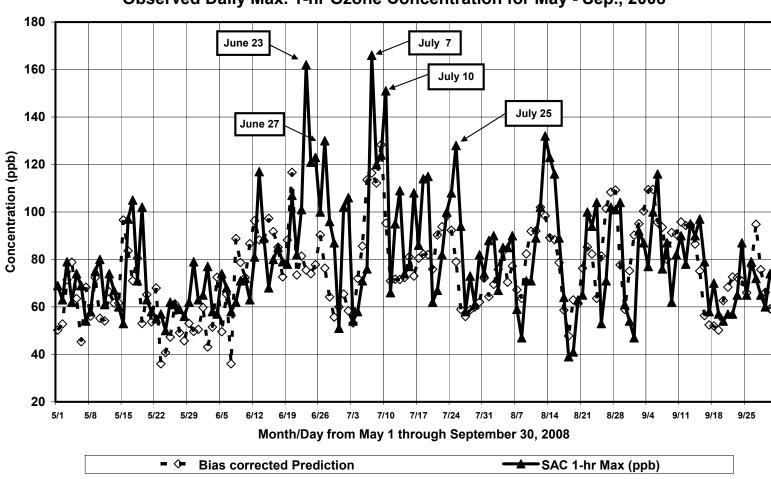
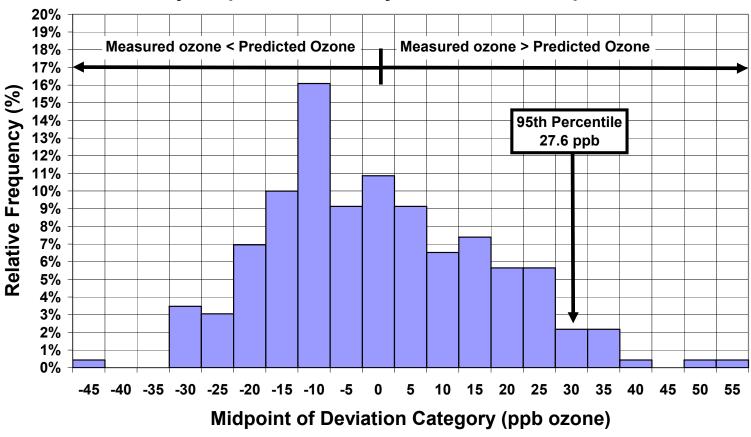


Figure 3. Distribution of Deviations: (Obs. Ozone - Pred. Ozone*)
Based on Data from Sacramento County, California in
May - Sept., 2007 with May 1 - June 22 and Sept., 2008



^{*} Predicted values are the daily predicted 1-hour maximum values forecasted for Sacramento County by Sonoma Technologies, Inc. using forecasted meteorology from

Figure 4. Comparison of May 2007 and May 2008 Regarding Observed - Predicted Daily Max Ozone

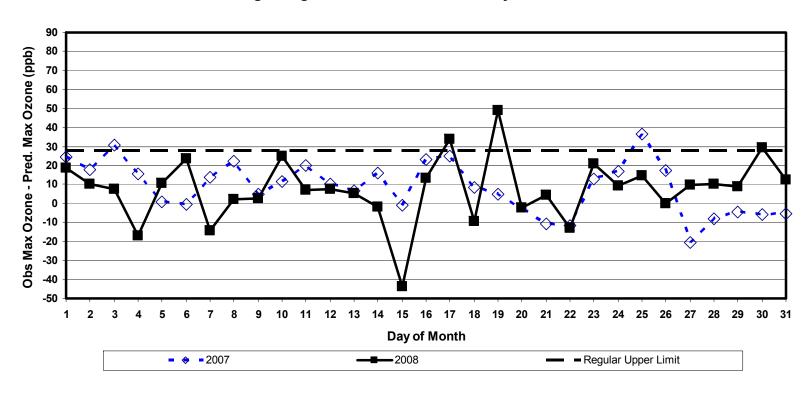


Figure 5. Comparison of June 2007 and June 2008 Regarding Observed - Predicted Daily Max Ozone

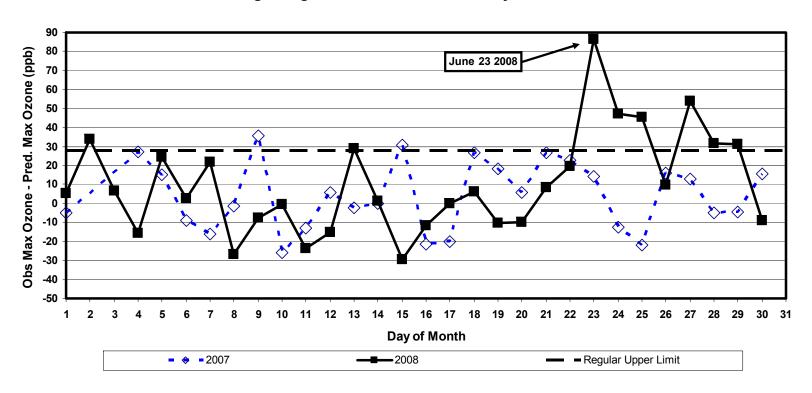


Figure 6. Comparison of July 2007 and July 2008 Regarding Observed - Predicted Daily Max Ozone

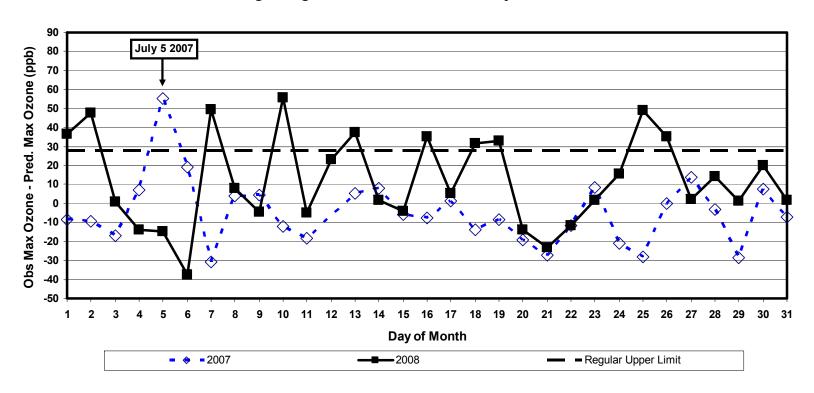


Figure 7. Comparison of August 2007 and August 2008 Regarding Observed - Predicted Daily Max Ozone

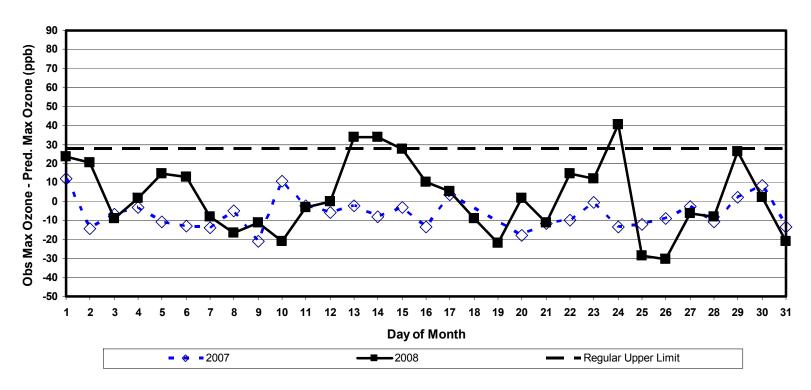


Figure 8. Comparison of September 2007 and September 2008 Regarding Observed - Predicted Daily Max Ozone

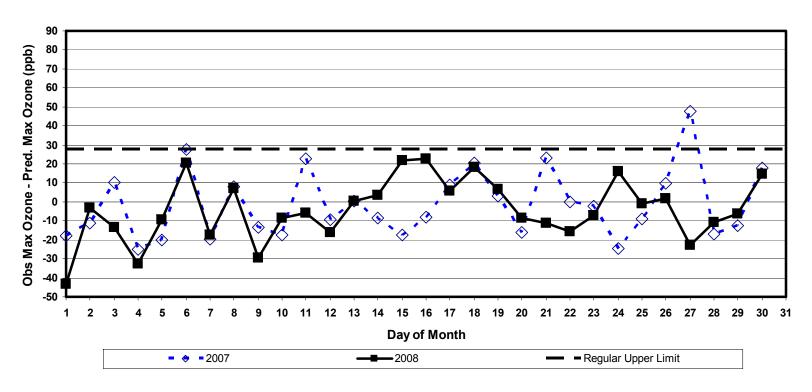


Figure 9. Comparison of October 2007 and October 2008 Regarding Observed - Predicted Daily Max Ozone

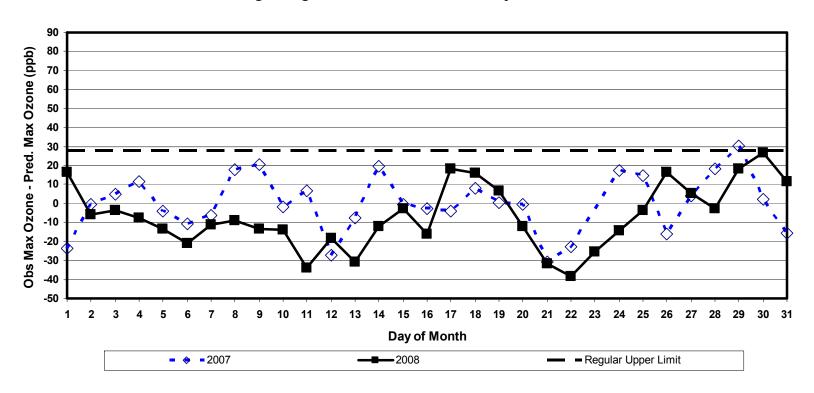


Figure 10. Sacramento County: Possible Exceptional Events
Based on May 2008 Daily Max 1-Hour Ozone

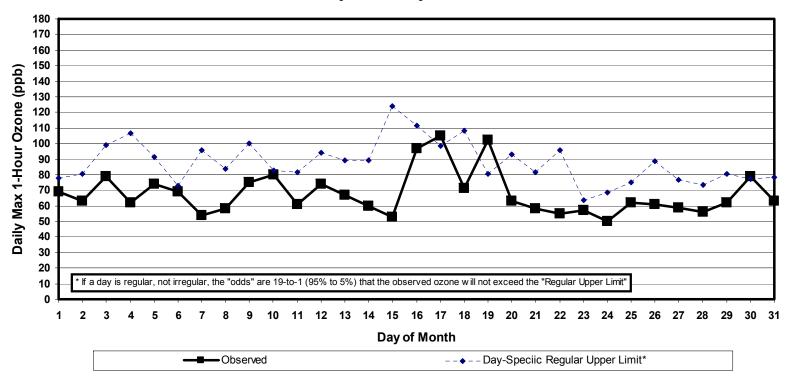


Figure 11. Sacramento County: Possible Exceptional Events
Based on June 2008 Daily Max 1-Hour Ozone

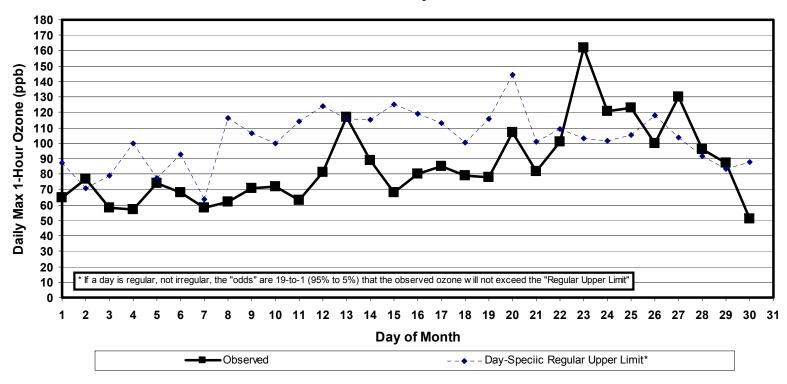


Figure 12. Sacramento County: Possible Exceptional Events
Based on July 2008 Daily Max 1-Hour Ozone

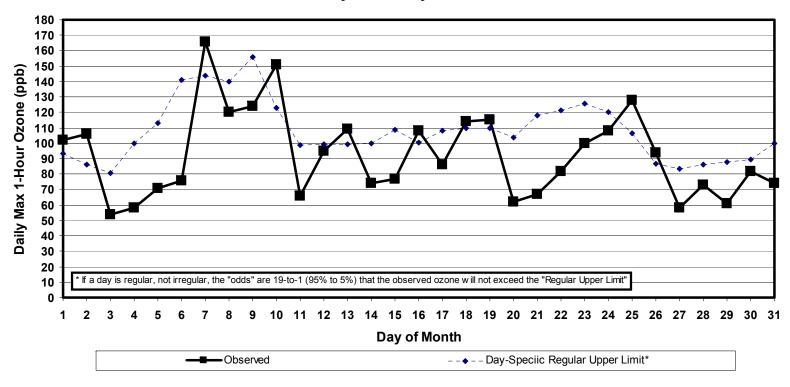


Figure 13. Sacramento County: Possible Exceptional Events
Based on August 2008 Daily Max 1-Hour Ozone

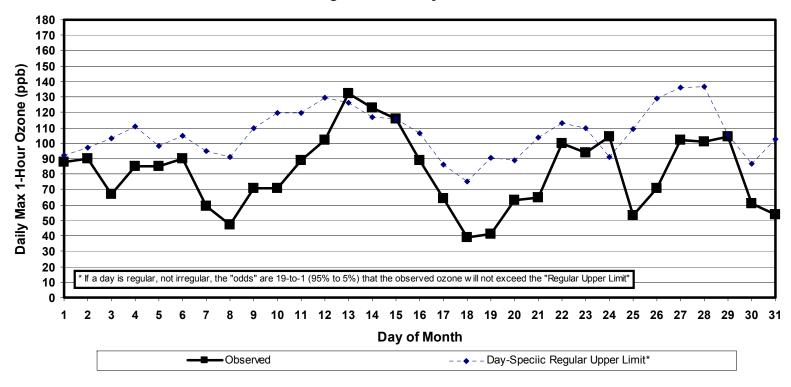


Figure 14. Sacramento County : Possible Exceptional Events Based on September 2008 Daily Max 1-Hour Ozone

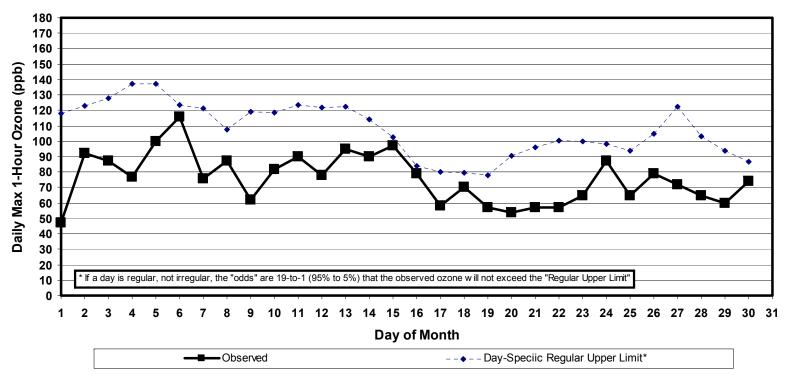


Table 1. Exceptional Event Data Summary					
		Regular			
	Observed	Upper	Exclude		
Date	Ozone	Limit *	from D.V. ?		
23-Jun-2008	162 ppb	103 ppb	yes		
27-Jun-2008	130 ppb	104 ppb	yes		
7-Jul-2008	166 ppb	144 ppb	no		
10-Jul-2008	151 ppb	123 ppb	yes		
25-Jul-2008	128 ppb	107 ppb	yes		

^{*} This "Regular Upper Limit" is day-specific, and the "odds" are 19 to 1 (95% to 5%) against a regular day exceeding its day-specific limit. So, the 123 ppm limit on July 10, 2008, is already a hard-to-reach limit, providing strong evidence that the day would not exceed the Federal 1-hour standard if it were a regular, not an irregular, day.